CS281A/Stat241A: Statistical Learning Theory

Lecture 20 (11/4/02)

Lecturer: Michael I. Jordan *Scribes: Meingast and Schonberg*

1 Hidden Markov Models

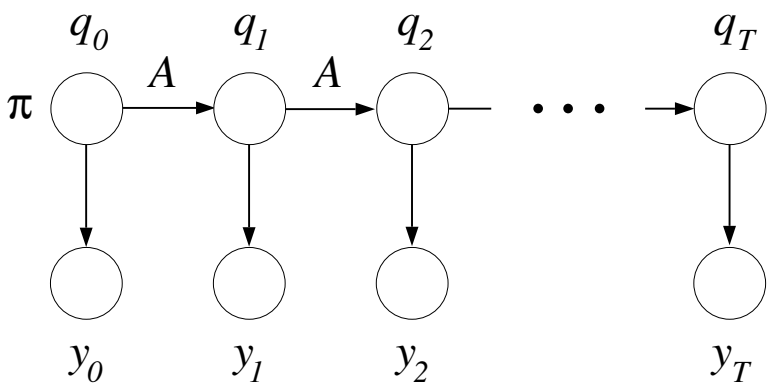


Figure 1: The representation of a HMM as a graphical model. Each vertical slice represents a time step. The top node in each slice represents the multinomial q_t variable and the bottom node in each slice represents the observable y_t variable.

Applying EM Algorithm

The E-Step is:

$$\begin{aligned}
\alpha(q_{t+1}) &= \sum_{q_t} \alpha(q_t) a_{q_t,q_{t+1}} p(y_{t+1}|q_{t+1}) \\
\beta(q_t) &= \sum_{q_{t+1}} \beta(q_{t+1}) a_{q_t,q_{t+1}} p(y_{t+1}|q_{t+1}) \\
\gamma(q_t) &= \frac{\alpha(q_t)\beta(q_t)}{p(y)}
\end{aligned}$$

where,

$$\begin{aligned}
\alpha(q_t) &= p(y_0, \ldots, y_t, q_t) \\
\beta(q_t) &= p(y_{t+1}, \ldots, y_T|q_t)
\end{aligned}$$

M-Step corresponds to using "soft counts" to estimate the parameters.

Three inference problems:

- estimation problem : an application of the E-Step
- computing the likelihood : run $\alpha(q_t)$ up to last time step T

$$\begin{aligned} \alpha(q_t) &= p(y_1, y_2, ..., y_T, q_t) \\ \sum_{q_t} \alpha(q_t) &= p(y_1, y_2, ..., y_T) \end{aligned}$$

- segmentation
  - goal: to determine most highly probable state sequence
  - use max-product algorithm: $\max_q P(q, y|\theta)$
  - in the HMM case, the max-product algorithm is called the Viterbi algorithm
- Numerical Issues (underflow):
  - if an HMM is implemented and it is tried on a sequence of approximately 10, a segmentation fault may occur due to underflow.
  - Need to scale α and β and keep track of the scaling to correctly compute the likelihood.

In the end of the coming discussions, this table should be fully understood:

	Static	Dynamic
Multinomial (Discrete)	Finite Mixture Models	Hidden Markov Models
Gaussian (Continuous)	Factor Analysis	Kalman Filtering

2 Jointly Gaussian Random Variables

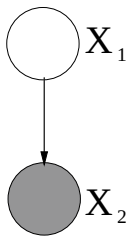


Figure 2: Jointly Gaussian Random Variables.

For random vector X_1 of length p and X_2 of length q where

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

We can write a joint Gaussian distribution for x_1 and x_2 using these partitioned parameters:

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{(p+q)/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right\}$$

Using Equation 2, the goal is then to find $p(x_2)$ and $p(x_1|x_2)$.
Due to closure, we know that both the marginals and the conditional densities will be Gaussian.

2.1 Partitioned Matrices

Consider a general partitioned matrix:

$$M = \begin{bmatrix} E & F \\ G & H \end{bmatrix},$$

where we assume that both E and H are invertible. Performing "Gauss-Jordan" Elimination of this matrix on blocks gives

$$\begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} = \begin{bmatrix} E - FH^{-1}G & 0 \\ 0 & H \end{bmatrix}.$$

We define the *Schur complement* of the matrix M with respect to H, denoted M/H, as the term $E - FH^{-1}G$ that appears in the block diagonal matrix.
It can be shown that $|M| = |M/H||H|$, which is a motivation for the notation used.
Recalling that $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$ and using the form above, we can derive the following.

$$\begin{bmatrix} E & F \\ G & H \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} \begin{bmatrix} (M/H)^{-1} & 0 \\ 0 & H^{-1} \end{bmatrix} \begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix} \quad (1)$$

$$= \begin{bmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{bmatrix}, \quad (2)$$

Further note that we could have done the Schur compliment with the opposite bias.

$$\begin{bmatrix} E & F \\ G & H \end{bmatrix}^{-1} = \begin{bmatrix} E^{-1} + E^{-1}F(M/E)^{-1}GE^{-1} & -E^{-1}F(M/E)^{-1} \\ -(M/E)^{-1}GE^{-1} & (M/E)^{-1} \end{bmatrix}.$$

Both of these expressions must be equal, thus

$$(M/H)^{-1} = E^{-1} + E^{-1}F(M/E)^{-1}GE^{-1}$$

This is known as the matrix inversion lemma. It is given in its more common form below.

$$(E - FH^{-1}G)^{-1} = E^{-1} - E^{-1}F(H - GE^{-1}F)^{-1}GE^{-1}$$

It is an important formula, and can further be expressed as below.

$$-(M/H)^{-1}FH^{-1} = -E^{-1}F(M/E)^{-1}$$

2.2 Marginalization and Conditioning

The goal of the previous section was to invert the covariance matrix of the multivariate Gaussian. Using this work to split the joint distribution Eq. 2 into a marginal probability for x_2 and a conditional probability for x_1 according to the factorization $p(x_1, x_2) = p(x_1|x_2)p(x_2)$. Focusing first on the exponential factor

$$\exp\left\{-\frac{1}{2}\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}\right\}$$

$$= \exp\left\{-\frac{1}{2}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}^T \begin{bmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{bmatrix} \begin{bmatrix} (\Sigma/\Sigma_{22})^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{bmatrix} \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}\right\}$$

$$= \exp\left\{-\frac{1}{2}(x_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2))^T(\Sigma/\Sigma_{22})^{-1}(x_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2))\right\}$$

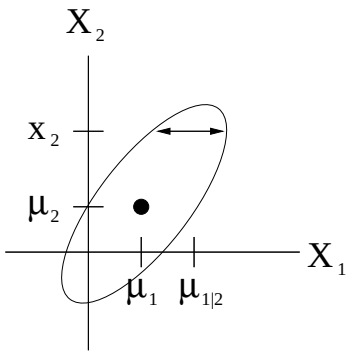
$$\cdot \exp\left\{-\frac{1}{2}(x_2-\mu_2)^T\Sigma_{22}^{-1}(x_2-\mu_2)\right\}. \quad (3)$$

Noticing that this is already factorized, we can arrive at the marginal and condition distributions.

$$X_2 \sim \mathcal{N}(\mu_2, \Sigma_{22})$$

$$X_1|x_2 \sim \mathcal{N}(\mu 1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2), \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

The conditional density can be seen as the original variance Σ_{11} of X_1 minus the loss in uncertainty $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ provided for by the knowledge of X_2.



3 Factor Analysis (FA)

- "Probablistic version of PCA (Principal Component Analysis)"
- Gaussian analog of mixture model

	Deterministic	Probablistic
Multinomial	K-means	Mixture Models
Gaussian	PCA	FA

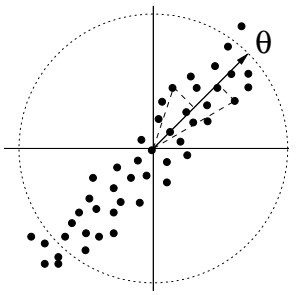


Figure 3: Assume the data set is zero mean. $\{y_n\}_{n=1}^N$ data vectors

The goal is to find a vector θ such that the projection of the data on θ have maximal variance. Give θ unit norm.

$$\max_\theta \frac{1}{N}\sum_{n=1}^{N}(\theta^T y_n)^2 = \max_\theta \frac{1}{N}\sum_{n=1}^{N}(z_n)^2$$

This is the empirical variance of z_n with the constraint that $\theta^T\theta = 1$.
Use Lagrangian

$$\begin{aligned}
\tilde{J} &= \frac{1}{N}\sum_n(\theta^T y_n)^2 - \lambda\theta^T\theta && (4)\\
&= \frac{1}{N}\sum_n(y_n^T\theta)^2 - \lambda\theta^T\theta && (5)\\
&= \frac{1}{N}\sum_n(\theta^T y_n y_n^T\theta) - \lambda\theta^T\theta && (6)\\
&= \theta^T(\frac{1}{N}\sum_n y_n y_n^T)\theta - \lambda\theta^T\theta && (7)\\
&= \theta^T S\theta - \lambda\theta^T\theta && (8)
\end{aligned}$$

where $S = \frac{1}{N}\sum_n y_n y_n^T$.
Maximize this with respect to θ

$$\begin{aligned}
\frac{\partial\tilde{J}}{\partial\theta} &= 0\\
\Rightarrow &= 2S\theta - 2\lambda\theta = 0\\
&\Rightarrow S\theta = \lambda\theta
\end{aligned}$$

i.e. θ is an eigenvector of S and λ is its eigenvalue.
FA won't get fooled by directed noise.